Starter Questions

Complete the exam question hand out.

You may need the following formula:

Standard deviation

$$\sqrt{\frac{\Sigma(x-\bar{x})^2}{n}} = \sqrt{\frac{\Sigma x^2}{n} - \bar{x}^2}$$

Starter Questions

11 Estimate the standard deviation of the times given in this frequency table.

Time (minutes)	Frequency
$5 \leq t < 10$	4
$10 \leq t < 20$	2
$20 \leq t < 25$	6
$25 \leq t < 40$	1

Circle your answer.

[1 mark]

7.3 7.8 8.5 9.2

Starter Questions

14 Given that $\sum x = 364$, $\sum x^2 = 19412$, $n = 10$, find σ, the standard deviation of X.

Circle your answer.

[1 mark]

24.8 44.1 616.2 1941.2

om the formula sheet:

Standard deviation

$$\sqrt{\frac{\Sigma(x-\bar{x})^2}{n}} = \sqrt{\frac{\Sigma x^2}{n} - \bar{x}^2}$$

$$\bar{x} = \frac{\sum x}{n} = \frac{364}{10} = 36.4$$

$$\sigma = \sqrt{\frac{19412}{10} - 36.4^2}$$

Starter Questions

15 A random sample of ten CO_2 emissions was selected from the Large Data Set.

The emissions in grams per kilogram were:

13 45 45 0 49 77 49 49 49 78

15 (a) Find the standard deviation of the sample.

[1 mark]

From the calculator:

Starter Questions

15 (b) An environmentalist calculated the average CO_2 emissions for cars in the Large Data Set registered in 2002 and in 2016.

The averages are listed below.

Year of registration	2002	2016
Average CO_2 emission	171.2	120.4

The environmentalist claims that the average CO_2 emissions for 2002 and 2016 combined is 145.8

Determine whether this claim is correct.

Fully justify your answer.

[2 marks]

There would have to be the same number of cars registered in 2002 as there are in 2016 for this to be true. There are more cars registered in 2016 than

L Data presentation and interpretation

L4

Recognise and interpret possible outliers in data sets and statistical diagrams.

Select or critique data presentation techniques in the context of a statistical problem.

Be able to clean data, including dealing with missing data, errors and outliers.

Teaching guidance

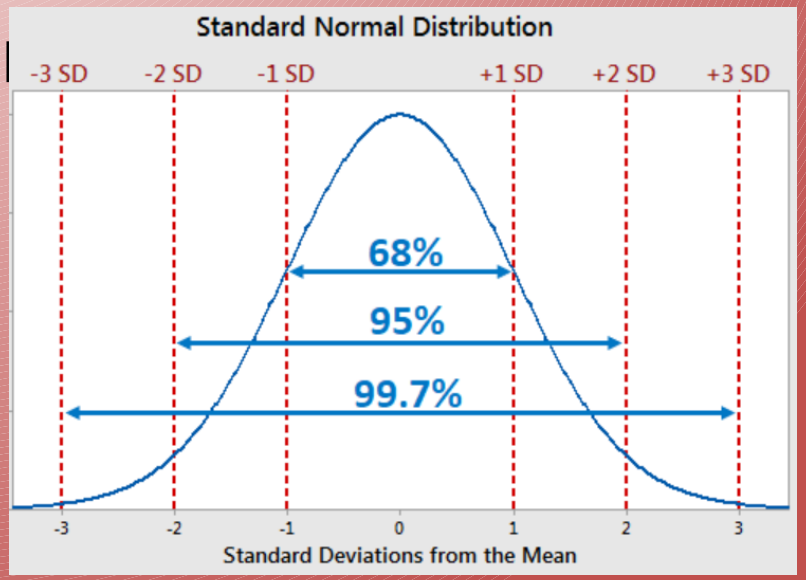
Students should be able to:

- identify outliers either from a given rule or from observation of a given diagram
- comment on the likely effect of removing the outlier
- identify clear errors in data and comment on or suggest subsequent actions needed
- select which of the representations in sections L1 and L2 is appropriate for representing given data sets
- criticise, in context, a given representation.

9.2 Central Tendency & Spread

Approximate Properties of the Standard Deviation

- About two thirds of values lie within one standard deviation of the mean
- About 95% of values lie within two standard deviations of the mean
- Almost all values lie within three standard deviations of the mean



9.2 Central Tendency & Spread

Outliers

Outliers are values that lie significantly outside the typical set of values of the variable.

An outlier can be defined in several ways.

For example, any value that lies outside the interval:

This is not the only rule, you will be told which rule to apply if another is needed.

Outliers may indicate natural variation in the data or may be the result of measurement and recording errors. If an outlier is due to an error,

9.2 Central Tendency & Spread

Example 1a

Define an outlier as a value *more than two standard deviations from the mean*.

A group of 15 students complete a timed test for their homework. Their times (in minutes) are recorded:
32, 34, 33, 37, 39, 39, 42, 45, 41, 40, 40, 44, 13, 36, 36

a) Calculate the mode, median and mean of the data

Mode: 36, 39 and 40

9.2 Central Tendency & Spread

Example 1b

Define an outlier as a value *more than two standard deviations from the mean*.

A group of 15 students complete a timed test for their homework. Their times (in minutes) are recorded:
32, 34, 33, 37, 39, 39, 42, 45, 41, 40, 40, 44, 13, 36, 36

b) Show that there is exactly one outlier in the data

Mean: = 36.7

13 is the only value outside this range.

9.2 Central Tendency & Spread

Example 1ci

Define an outlier as a value *more than two standard deviations from the mean*.

A group of 15 students complete a timed test for their homework. Their times (in minutes) are recorded:
32, 34, 33, 37, 39, 39, 42, 45, 41, 40, 40, 44, 13, 36, 36

c) Give one possible reason for:

i. Removing the outlier

The outlier could be an error. It could have been recorded incorrectly, or a parent could have helped, meaning it is not a valid test result. In

9.2 Central Tendency & Spread

Example 1cii

Define an outlier as a value *more than two standard deviations from the mean*.

A group of 15 students complete a timed test for their homework. Their times (in minutes) are recorded:
32, 34, 33, 37, 39, 39, 42, 45, 41, 40, 40, 44, 13, 36, 36

c) Give one possible reason for:

ii. Not removing the outlier

If this is a true value, removing it gives a false picture, underestimating the variation of the results.

9.2 Central Tendency & Spread

Example 1d

Define an outlier as a value *more than two standard deviations from the mean*.

A group of 15 students complete a timed test for their homework. Their times (in minutes) are recorded:
32, 34, 33, 37, 39, 39, 42, 45, 41, 40, 40, 44, 13, 36, 36

d) A teacher investigates the outlier and decides to remove it. Without further calculation explain how removing this value would affect your answers to part a.

The mode and median are not affected, the mean increases.

9.2 Central Tendency & Spread

Comparing Measures of Central Tendency

Statistic	Advantages	Disadvantages
Mode	Useful for non-numeric data. Not usually affected by outliers. Not usually affected by errors or omissions. Is always an observed data point.	Doesn't use all the data. May not be representative if it has a low frequency. There may be more than one mode.

9.2 Central Tendency & Spread

Comparing Measures of Central Tendency

Statistic	Advantages	Disadvantages
Median	Not affected by outliers. Not significantly affected by errors.	Doesn't make use of all the data.
Mean	When the data set is very large, a few extreme values have negligible impact. Uses all the data.	Can be affected by outliers. When the data set is small, a few extreme values or errors have a

9.2 Central Tendency & Spread

Comparing Measures of Spread/Dispersion

Statistic	Advantages	Disadvantages
Range	Reflects the full set of data	Distorted by outliers
IQR	Not distorted by outliers	Does not reflect all the data
Standard Deviation	Uses all the data. When the data set is very large, a few outliers have negligible effect.	When the data set is very small, a few outliers have a big impact.

9.2 Central Tendency & Spread

Example 2a

A corner shop records the volume of mineral or spring water it sells every day for a week, to the nearest gallon:

3, 4, 6, 6, 1, 0, 25

a) Calculate an appropriate measure of spread. Explain your choice of statistic.

$Q_1 = 1$, $Q_3 = 6$, IQR = 6 – 1 = 5

The range and standard deviation are both distorted by the outlier 25 but the IQR is not significantly affected.

9.2 Central Tendency & Spread

Example 2b

A corner shop records the volume of mineral or spring water it sells every day for a week, to the nearest gallon:

3, 4, 6, 6, 1, 0, 25

b) Explain why the mode is not an appropriate measure of central tendency.

This data set is small with only two instances of the mode, and thus is not representative of the data. There is enough data to calculate the median, which is more representative of the data.